Contribution ID: **56** Type: **Lightning Talk**

Optimizing Sparse Neural Architectures for Low-Latency Anomaly Detection

Monday 25 September 2023 18:15 (5 minutes)

Within the framework of the L1 trigger's data filtering mechanism, ultra-fast autoencoders are instrumental in capturing new physics anomalies. Given the immense influx of data at the LHC, these networks must operate in real-time, making rapid decisions to sift through vast volumes of data. Meeting this demand for speed without sacrificing accuracy becomes essential, especially when considering the time-sensitive nature of identifying key physics events. With ultra low-latency requirements at the trigger, we can leverage hardware-aware neural architecture search techniques to find optimal models. Our approach leverages supernetworks to explore potential subnetworks through evolutionary search and unstructured neural network pruning, facilitating the discovery of high-performing sparse autoencoders. For efficient search, we train predictor networks for each objective, lowering the sample cost of evolutionary search. Here, we optimize for the post-pruning model. Due to the unique nature of reconstruction-based anomaly detection methods, we explore how neural network pruning and sparsity affect the generalizability on out-of-distribution data in this setting.

Primary authors: DEMLER, Dmitri; WEITZ, Jason (UC San Diego); DUARTE, Javier Mauricio (Univ. of California San Diego (US)); MCDERMOTT, Luke (UC San Diego & Modern Intelligence); TRAN, Nhan (Fermi National Accelerator Lab. (US))

Presenter: MCDERMOTT, Luke (UC San Diego & Modern Intelligence)

Session Classification: Contributed Talks

Track Classification: Contributed Talks